AI-Assisted Safety Assessment: Automation Bias as an Emergent Risk

Adam Stein[a]
[a] Breakthrough Institute, Washington D.C., United States, adam@thebreakthrough.org

Abstract: Automation bias introduced from the use of AI in safety and regulatory applications is a human-organizational reliability problem, rather than a failure of AI models. Evidence across aviation, medicine, and nuclear contexts shows that even highly accurate systems systematically shift human behavior: experts reduce independent verification, narrow hypothesis generation, and become less likely to detect errors when automation fails.

This paper treats automation bias as an emergent risk factor within AI-enabled risk assessment processes. In probabilistic safety assessment and regulatory review, the relevant question is not only model accuracy, but how AI outputs alter vigilance, accountability, and error detection over time. Slow, document-intensive regulatory contexts differ from time-critical operational environments, but they introduce their own risks: habituation, gradual skill atrophy, and institutionalized over-reliance.

Drawing on cross-industry evidence and existing nuclear human-factors frameworks, the paper outlines practical design and governance strategies for AI-assisted probabilistic risk assessment and safety review. These include structured independent verification, explicit representation of model uncertainty and limits, traceability to underlying evidence, periodic manual review exercises, and organizational accountability mechanisms that prevent diffusion of responsibility. The objective is to ensure that AI integration strengthens decision-relevant risk assessment capability without degrading the human judgment that underpins safety decisions.

1. INTRODUCTION

The nuclear safety community has engaged seriously with the question of how artificial intelligence (AI) systems can improve probabilistic safety assessment and regulatory review. Much of this discussion centers on model performance: accuracy rates, false-positive thresholds, coverage completeness, and validation against historical cases. That framing is necessary, but it is insufficient. It treats AI integration primarily as an engineering problem with a technical solution, when the harder and less-examined problem is how sustained human reliance on AI systems changes the character of review itself.

Automation bias is not a rare failure mode or a problem confined to poorly designed systems. It is a well-documented behavioral phenomenon that affects expert practitioners across domains, operates even when automated systems are highly accurate, and can intensify as repeated reliability builds trust [1,2]. When a nuclear safety reviewer, a probabilistic risk analysis (PRA) analyst, or a regulatory engineer routinely receives AI-generated outputs before forming independent judgments, the cognitive architecture of that review changes, not catastrophically or immediately, but gradually and in ways that are difficult to detect from within the process.

This paper focuses on AI-assisted review of risk assessment and safety evidence, not on the use of AI to construct PRA models. The concern is how AI-generated summaries, conformance checks, anomaly flags, evidence maps, and preliminary technical judgments may alter the vigilance, accountability, and error-detection behavior of reviewers over time. Large-scale deployment in nuclear regulatory review remains limited, and the relevant use cases are still emerging. This means direct operational evidence is necessarily limited. That scarcity, however, is no warrant for complacency. The contribution at this stage is anticipatory: identifying plausible human-organizational failure modes before they become embedded in routine practice.

Automation bias should be treated as an emergent risk factor in AI-assisted PRA and regulatory review: a systemic behavioral hazard that, if unaddressed, can degrade the quality of the human judgment that ultimately underpins safety decisions. Behavioral effects of AI assistance must be managed as deliberately as the technical performance of AI systems. These are different problems requiring different tools.

The paper proceeds as follows. Section 2 defines automation bias and its mechanisms. Section 3 reviews evidence from aviation, medicine, and nuclear-adjacent applications. Section 4 analyzes the distinct risk profile of slow, document-intensive regulatory review relative to the time-critical scenarios where automation bias has been most studied. Section 5 examines existing Nuclear Regulatory Commission (NRC) and International Atomic Energy Agency (IAEA) human-factors frameworks and identifies where current guidance requires extension for AI-assisted review. Section 6 presents practical design and governance recommendations. Section 7 concludes.

2. AUTOMATION BIAS: MECHANISMS AND SCOPE

Automation bias was first characterized in studies of human decision-making with automated aids and has since been examined across aviation, medicine, process control, and military command systems [1,3]. The term refers to the inappropriate reliance on automated outputs in ways that reduce independent information-seeking or error detection. Two error modes are especially important for AI-assisted review.

Table 1: Automation bias error modes and their relevance to AI-assisted safety review

Error mode	Mechanism	Relevance to AI-assisted safety review
Error of omission	The operator fails to act because the automated system did not issue a prompt or alert. Automation silence is treated as informative, even when it may reflect only detection limits or incomplete training data.	AI-generated conformance checks, summaries, or anomaly screens may cause reviewers to overlook issues that the system did not flag.
Error of commission	The operator follows incorrect automated guidance despite contradictory evidence available from other sources or from prior judgment. The automated output displaces, rather than supplements, independent analysis.	AI-generated findings may anchor reviewers toward an incorrect conclusion, especially when the output appears complete, authoritative, or consistent with prior experience.

Both error types share a common mechanism: automation functions as a heuristic replacement for vigilant information seeking, rather than as a supplementary tool subject to critical evaluation [3]. A systematic review described automation bias as a robust and generic effect across research fields, including cases where users were told that the system was imperfect and where contradictory evidence was available [1].

A cluster of related phenomena bears directly on slow regulatory work, because each traces how reliance compounds over time rather than within a single decision.

Table 2: Time-dependent reliance phenomena and their implications

Phenomenon	Mechanism	Implication for PRA and regulatory review
Automation complacency	Monitoring intensity declines when automated systems perform reliably, degrading detection of system malfunctions or missed cues [2,4].	Reviewers may stop actively sampling the underlying evidence when AI outputs have usually been correct.
Learned carelessness	High reliability can train users to accept outputs with minimal scrutiny because errors are rare during normal use [1,4].	A 95 percent accurate review aid may produce its most consequential failures after users have learned that checking is usually unnecessary.

Phenomenon	Mechanism	Implication for PRA and regulatory review
Hypothesis narrowing	Once an automated system frames the likely issue, users may search less broadly for alternative explanations.	AI-generated issue framing can reduce exploration of weak signals, edge cases, or alternative interpretations not emphasized by the system.

Accountability also affects reliance behavior. Experimental work on automated decision environments found that internalized personal accountability reduced both omission and commission errors, whereas procedural compliance alone did not provide the same protection [5]. In the context of regulatory review, if accountability is structurally diffused across teams, procedures, and AI systems, the protective effect of individual responsibility may disappear.

3. EVIDENCE FROM SAFETY-CRITICAL DOMAINS

3.1 Clinical Medicine: Quantified Performance Degradation

Clinical medicine provides the largest body of controlled research on automation bias effects in decision support settings. Two systematic reviews establish the foundational pattern: AI assistance can improve mean performance while simultaneously introducing errors that would otherwise not occur [1,6]. The net result may be positive on average, but the distribution of outcomes can conceal systematic harm for a predictable subset of decisions.

The performance benefits are real and measurable. In a large randomized clinical vignette study across 13 U.S. states, accurate AI model assistance increased clinicians' baseline diagnostic accuracy by 2.9 to 4.4 percentage points [7]. Computer-based assistants have surfaced diagnoses that experts had not originally considered but, on review, judged clinically relevant [8]. A recent randomized controlled trial in a chest pain triage scenario found similar improvement, with GPT-4 assistance raising accuracy from 47 percent to 65 percent in one patient group and from 63 percent to 80 percent in another, without introducing demographic biases [9].

The performance degradation from reliance on flawed AI is proportionally larger and more consequential. Across controlled clinical decision-support studies, negative consultation rates (cases where clinicians changed a correct pre-consultation answer to an incorrect one after receiving automated guidance) consistently range from 6 percent to 11 percent [1,10,11]. In ECG interpretation, residents' diagnostic accuracy fell from 57 percent to 48 percent when using a system whose outputs were deliberately manipulated to be inaccurate, providing a controlled demonstration of the commission-error mechanism [12]. In the randomized vignette study, systematically biased AI reduced clinician accuracy by 11.3 percentage points below baseline, more than three times the benefit produced by accurate AI [7].

Explanations do not rescue the situation. In the same vignette study, biased AI with explanations still left accuracy 9.1 percentage points below baseline, and the difference relative to biased AI without explanations was statistically insignificant [7]. An accompanying editorial emphasized the practical concern: even in controlled settings without ordinary clinical time pressure, clinicians deferred to AI over their own judgment when contradictory evidence was present [13]. Explainability can support appropriate reliance, but current forms of explanation do not reliably function as a safeguard against automation bias.

Expertise attenuates automation bias but does not eliminate it. Systematic reviews find that expert operators are generally less vulnerable than novices, but still remain subject to commission and omission errors under automation [1,6].

3.2 Commercial Aviation: Systemic Over-Reliance Under Pressure

Aviation was the first domain in which automation bias was systematically characterized, and it provides clear evidence of how automated aids can alter expert behavior under high-consequence conditions [2,14].

The Air France Flight 447 accident in 2009 illustrates automation mode confusion and its downstream consequences. When the autopilot disengaged following pitot icing, the crew made recovery decisions inconsistent with the manual recovery needed for the event. The final accident report identifies issues associated with automation, crew understanding, and manual recovery performance as part of the accident sequence [15]. This case should not be reduced to automation bias alone, but it illustrates a broader human-automation problem: skills and expectations shaped by automation can become fragile when automation no longer performs as expected.

High-fidelity cockpit simulations provide more direct evidence of the commission-error mechanism. In controlled studies, pilots followed incorrect automated recommendations, including recommendations that contradicted their prior assessments and other available evidence [14]. In the same research program, pilots who had previously stated that they would not shut down an engine based on a false automated alert nevertheless complied when placed in the automated simulation environment. The result demonstrates the gap between stated intention and actual behavior under automation.

Research on automation levels also illustrates the performance-monitoring tradeoff. Higher automation can improve flight performance and reduce mental workload while simultaneously decreasing vigilance to primary flight instruments [4]. Aviation training can reduce these effects, particularly when training exposes operators to deliberate automation failures rather than only warning them that failures are possible, but the bias is not eliminated [1].

3.3 High-Consequence Defense Automation: Isolating the Organizational Pathway

Direct empirical evidence on automation bias in AI-assisted nuclear regulatory review remains limited because these workflows are still emerging. The medicine and aviation cases above already establish the behavioral pattern; the value of the case below is different. It isolates the organizational pathway to automation bias—how training and institutional culture, rather than any individual lapse, can render human oversight nominal—which is the mechanism most relevant to a regulatory setting.

The Patriot missile fratricides during Operation Iraqi Freedom in 2003 provide a consequential case. In two separate incidents, Patriot operators engaged friendly aircraft based on automated threat classifications—a British Tornado and a U.S. Navy F/A-18—resulting in three deaths. A post-incident analysis identified automation misuse, specifically unwarranted trust in the system's target classification, as a primary contributing factor enabled by training practices that had not adequately emphasized independent verification of automated threat identification [16].

The case illustrates the failure mode: well-trained professionals, operating a high-reliability system, in conditions that made independent verification cognitively costly. The result was human oversight that was nominal rather than substantive. The same analysis notes that the underlying problem was systemic, rooted in development, training, and organizational culture rather than individual crew error alone.

Nuclear facility regulation operates on very different timescales. That distinction is central to the analysis in Section 4.

4. THE SLOW-DECISION CONTEXT: A DISTINCT RISK PROFILE

Most automation bias research has been conducted in time-critical operational settings: aviation emergencies measured in seconds, clinical decisions measured in minutes, and missile defense decisions measured in fractions of a minute [1]. Nuclear PRA and regulatory review operate on fundamentally different timescales: individual technical reviews may extend over months, and licensing proceedings may extend over years. The absence of acute time pressure changes the behavioral risk profile in both

favorable and unfavorable directions. Effective governance requires distinguishing risks that are mitigated by deliberation from risks that are merely displaced into slower organizational processes.

4.1 Structural Advantages of Deliberative Review

Time reduces some acute drivers of heuristic processing, but it does not eliminate behavioral risk. Time pressure, high workload, and irreversibility are classic conditions for reliance on fast, intuitive judgment. Experimental evidence on risky decision-making supports this distinction, but also shows why it should be stated carefully. One set of experiments found that time pressure and time delay alter risk-taking behavior and that time-delay conditions reduce measurement noise relative to faster decision conditions [17]. A second set of experiments found that time pressure had little effect on pure gain decisions but affected loss and mixed-prospect decisions, including greater sensitivity to framing and aspiration-level heuristics [18]. These studies support that deliberative regulatory review is less exposed to the acute timing pressures that can amplify heuristic processing, but its protection depends on preserved independent reasoning, not on time alone.

Multiple review layers create opportunities for distributed verification. Nuclear regulatory processes typically involve preliminary analysis, staff technical review, management oversight, independent verification for novel technical issues, and, in some cases, public comment or Commission-level approval. NUREG-0711 provides one example of this structured review logic in the human-factors engineering context [19]. Each layer can catch errors that propagated from prior stages, including AI errors that survived earlier review. Time-critical operational settings afford no comparable redundancy.

Documentation requirements externalize reasoning and can protect review quality. Written justification for regulatory determinations forces engagement with the evidence base, creates an artifact that can be challenged and audited, and promotes deliberative rather than intuitive cognitive processing. That protection depends on whether documentation remains substantive. A review record that merely reproduces AI-generated reasoning without independent engagement would preserve the form of review while weakening its function.

4.2 Distinct Risks in the Slow-Decision Context

Slow, document-intensive review introduces failure modes that are largely absent from the fast-decision literature. The concern is gradual adaptation: reviewers, teams, and institutions may learn over time that AI outputs are usually reliable, that independent verification is costly, and that responsibility for checking is distributed across the process. Outright unreliability would prompt the institution to abandon the tool; the danger lies in outputs reliable enough to be trusted.

Table 3: Distinct automation-bias risks in the slow-decision regulatory context

Slow-context risk	Mechanism	Consequence for AI-assisted review
Habituation over extended time	Familiarity with a consistently reliable system produces desensitization on a timescale of months or years rather than minutes or experimental sessions [1,4].	Reviewers may come to treat AI silence as meaningful evidence that no issue exists.
Cognitive offloading of dense analyses	Nuclear safety review involves large, technical submissions that are difficult to comprehend in full. AI tools that handle comprehensiveness checks create strong incentives to accept outputs rather than verify them [1,19].	The apparent efficiency of AI assistance may reduce independent engagement with the underlying record.
Organizational skill atrophy	Skills that are not exercised degrade. If AI performs routine conformance checks or preliminary analysis, staff may lose fluency in the manual review practices needed to evaluate AI failures [4].	The organization may preserve review throughput while losing the capability needed for independent error detection.

Slow-context risk	Mechanism	Consequence for AI-assisted review
Diffusion of responsibility	Team review does not automatically mitigate automation bias; experimental evidence found that teams were no more successful than individuals at responding to irregularities that automation failed to flag [5].	Each reviewer may assume that someone else has checked the AI output, resulting in no substantive verification.

None of this equates slow review with emergency decision-making. The point is narrower: deliberation buffers some forms of automation bias even as it opens a different risk profile. The most important slow-context hazards are habituation, cognitive offloading, gradual skill erosion, and institutionalized over-reliance.

4.3 The Central Risk Question

For PRA and regulatory review, the operative risk question is not: *Will the AI make an error?* It is: *When the AI makes an error, will anyone catch it?*

That question points directly to the human-organizational factors that determine whether AI integration strengthens or degrades the overall review process: who has the skills, the mandate, and the cognitive engagement to detect AI errors; whether the accountability structure produces active rather than passive oversight; and whether the information environment is designed to support independent judgment or to anchor it. These are design and governance problems, not technical problems alone.

5. EXISTING FRAMEWORKS AND AREAS FOR EXTENSION

The NRC and IAEA have developed substantial human-factors guidance directly relevant to AI-assisted review. That guidance, however, points outward: it disciplines how the regulator evaluates the human-automation systems of applicants and licensees, not how the regulator conducts its own review. The extension this paper calls for is to redirect that established discipline inward, onto AI-assisted regulatory review.

NRC-sponsored research on emerging nuclear technologies provides an additional foundation for this extension. Prior work on emerging technologies in nuclear power plants identified automation, human-system integration, and other new technologies as sources of human factors engineering challenges that may require additional review attention [20]. Related work on small modular reactors identified human-performance issues associated with novel concepts of operation, staffing, automation, and human-system interfaces [21]. These reports do not address AI-assisted regulatory review directly. Their relevance is that they establish a nuclear human-factors precedent for treating changes in technology and concepts of operation as changes in the human role that should be considered during safety review, not after deployment.

5.1 Strengths of the Existing Foundation

Existing nuclear human-factors frameworks already contain concepts that can be adapted to AI-assisted PRA and safety review. The most relevant contributions concern function allocation, interface design, organizational readiness, workforce development, explainability, and graded deployment.

Table 4: Existing nuclear human-factors frameworks and their relevance to automation bias

Framework	Relevant contribution	Relevance to automation bias
NUREG-0711	Establishes a structured function-allocation framework requiring explicit justification for automation decisions, evaluation of monitoring responsibilities, and demonstration that operators	The same discipline can be applied to AI deployment in regulatory review: what should AI do, what must humans continue to do, and how is the oversight burden justified?

Framework	Relevant contribution	Relevance to automation bias
	can assume manual control when automation fails [19].	
NUREG-0700	Provides human-system interface guidance that can be applied to reliance-sensitive design issues, including salience, active engagement, and presentation of uncertainty information [22].	Interface design can either reinforce AI authority or preserve independent review.
NUREG-2261	Recognizes that AI integration requires regulatory readiness, organizational frameworks, partnerships, workforce development, and use-case development [23].	AI-system literacy and workforce development are necessary because one-time awareness training is insufficient to manage automation bias.
IAEA 2025 AI guidance	Addresses human-factors safety considerations, explainability requirements, and graded risk-informed approaches to AI deployment [24].	AI systems whose reasoning basis cannot be examined are more difficult to use safely in review settings because users cannot readily calibrate reliance or identify potential failure modes.

The value of these frameworks is that they treat automation as a human-system design problem and not merely a software-performance one. That orientation is the right starting point for AI-assisted review.

5.2 Areas Where Existing Guidance Requires Extension

The main limitation of existing guidance is that AI-assisted regulatory review creates a slow, distributed, document-intensive use case that existing operational guidance does not fully address.

Table 5: Areas where existing guidance requires extension for AI-assisted regulatory review

Area for extension	Why it matters for AI-assisted regulatory review
Slow-decision context specificity	Operational guidance tends to address automation bias where risk materializes in real time. Regulatory review creates a different risk profile involving habituation, skill atrophy, and responsibility diffusion over months-long timescales.
Long-term behavioral monitoring	Technical validation at deployment does not measure how human oversight changes after repeated exposure to reliable AI assistance. The learned-carelessness mechanism implies that oversight quality at deployment may not predict oversight quality after two years of use.
Team accountability design	Team review does not automatically prevent automation bias and may intensify responsibility diffusion if individual verification roles are unclear [5]. Existing human and organizational factors guidance recognizes team dynamics as safety-relevant, but does not yet specify how those dynamics should be managed when AI assists review [25].
Internalized versus procedural accountability	Evidence suggests that internalized accountability is more protective than external compliance alone [5]. AI-assisted review processes need accountability structures that preserve personal responsibility for technical judgment rather than distributing responsibility across the tool, the team, and the procedure.

These extensions are achievable. They do not require treating AI as uniquely unmanageable, but do require applying existing nuclear human-factors discipline to the distinct behavioral dynamics introduced by AI-assisted review.

6. DESIGN AND GOVERNANCE RECOMMENDATIONS

The following recommendations are grounded in the cross-industry evidence and directed specifically at the slow, deliberative context of PRA and nuclear regulatory review. They are organized around the five controls identified in the abstract: independent verification, uncertainty representation, evidence traceability, manual review exercises, and accountability mechanisms. Each control addresses a specific pathway by which automation bias can degrade review quality.

Table 6: Governance controls and the automation-bias mechanisms they address

Governance control	Mechanism addressed	Implementation focus
Structured independent verification	Anchoring, commission errors, and loss of independent judgment	Require safety-significant AI-generated conclusions to be checked by a qualified reviewer who forms an initial assessment before seeing the AI output.
Explicit representation of uncertainty and limits	Overconfidence in AI outputs and false precision	Present confidence, assumptions, sensitivity to inputs, known limits, and threshold proximity as first-class outputs rather than footnotes.
Traceability to underlying evidence	Passive acceptance and inability to audit reasoning	Link every AI-generated finding to specific regulatory requirements, source documents, technical standards, and factual claims.
Periodic manual review exercises	Skill atrophy and loss of AI-off capability	Conduct selected review tasks without AI assistance to preserve manual competence and calibrate the quality of AI-supported review.
Accountability mechanisms and behavioral monitoring	Responsibility diffusion, habituation, and learned carelessness	Assign named responsibility for review judgments and track behavior such as acceptance rates, disagreement rates, verification depth, and blind error-detection performance.

AI outputs should be framed as inputs to human judgment, not as conclusions awaiting confirmation. Interface design should visually and linguistically position AI-generated findings as one source of evidence among several. Reducing the salience of automated recommendations relative to the underlying evidence base supports independent evaluation, especially when paired with explicit uncertainty information and traceable source links [1,22].

Uncertainty and system limits should be prominent enough to shape review behavior. AI systems should not present single-point conclusions when the underlying evidence depends on assumptions, incomplete records, disputed interpretations, or cases near decision thresholds. A reviewer who sees a conformance conclusion accompanied by its assumptions and sensitivity to those assumptions is working in a different information environment than one who sees the conclusion alone. Explainability is relevant because users who cannot understand the basis for a conclusion cannot calibrate their reliance on it [3,24].

Traceability is the practical bridge between AI assistance and independent review. Every AI-generated finding should be traceable to the regulatory requirement, technical standard, source document, or factual claim on which it depends. This traceability enables verification, supports auditability, and allows organizations to identify common-cause AI errors that would remain invisible if conclusions were reviewed only one at a time.

Structured independent verification should be targeted to safety-significant findings. High-consequence AI-generated conclusions, especially findings that directly affect a safety case determination, should be reviewed by a qualified person who has not been exposed to the AI output before forming an initial assessment. This requirement should not be universal across all AI-assisted activities; it should apply where the consequence of an undetected AI error justifies the cost of independent review.

Manual review exercises should be treated as competency maintenance, not as symbolic resistance to AI. On a defined schedule, selected review activities should be conducted without AI assistance so that staff preserve the skills needed to evaluate AI outputs with genuine technical authority. The regulatory analog to deliberate automation-failure training in aviation is periodic practice under conditions where AI assistance is unavailable or intentionally withheld. Maintaining an AI-off capability also requires that underlying data, traditional analysis tools, and manual review procedures remain available.

Accountability structures should align responsibility with decision authority. The individual who reviews and approves an AI-generated finding should bear clear responsibility for that finding's correctness. Accountability should not be diffused to the review team, the procedure, or the AI system's documented historical accuracy. Because internalized accountability can reduce automation-bias errors, sign-off structures should make clear who has exercised independent judgment and on what basis [5].

Organizational culture must protect independent judgment when it conflicts with AI outputs. If schedule pressure, implicit reputational costs, or managerial friction penalize reviewers who challenge AI-generated findings, formal mandates for independent review will not be sufficient. Review processes should make disagreement visible and legitimate by requiring reviewers to document doubts, alternative interpretations considered, and cases where AI outputs diverged from expert judgment.

AI-system literacy should be treated as domain-specific technical knowledge. Staff interacting with AI-assisted review systems should understand the system's training basis, the data from which its judgments derive, the issue categories where reliability degrades, the conditions under which it is likely to err, and the mechanisms of automation bias specific to their review context. This is analogous to understanding the limitations of a safety-analysis code, and it should be reflected in qualification requirements. NUREG-2261 moves in this direction; AI-assisted review should make the requirement explicit [23].

Behavioral monitoring should accompany technical performance monitoring. System-level monitoring should track the rate at which AI recommendations are accepted or substantively modified, the frequency and quality of documented disagreements, the depth of independent verification actually performed, and detection rates for intentionally inserted errors in periodic blind testing. These metrics characterize the quality of human oversight rather than the accuracy of the AI system. Declining modification rates or declining blind-test detection rates may be leading indicators of accumulating automation bias that ordinary technical performance metrics will not capture.

Independent audits should evaluate the complete human-AI review process, not just AI outputs. Periodic audits should examine whether accountability structures are functioning as intended, whether skill-maintenance exercises are being conducted and evaluated, whether disagreement documentation is substantive rather than formulaic, and whether behavioral indicators show trends consistent with habituation or skill degradation. Technical audits of AI system performance remain necessary, but organizational audits of the human oversight process are a distinct requirement.

7. CONCLUSIONS

AI integration into PRA and nuclear regulatory review offers distinct performance advantages: consistency, comprehensive coverage of large document sets, speed, and accessibility of accumulated regulatory knowledge. These advantages are real and worth pursuing. The risk this paper identifies is not that AI will be used, but that it will be used in ways that systematically erode the human judgment it is intended to support. Evidence from aviation, medicine, and nuclear-adjacent applications converges on one finding too easily discounted: sustained reliance on accurate automated systems reshapes human behavior in predictable, consequential ways. Automation bias is not a rare failure mode or a symptom of inadequate training; it is a recurring behavioral tendency under reliable automation, and it should be treated as a human-organizational hazard in AI-assisted safety review.

The slow, deliberative context of nuclear regulatory review provides structural advantages over the crisis scenarios that dominate the automation-bias literature. Multiple review layers, documentation requirements, and the absence of acute time pressure are notable mitigants. Experimental decision research suggests that timing conditions can alter risk behavior, measurement noise, and sensitivity to framing or aspiration-level heuristics [17,18]. But the absence of acute time pressure is not a complete safeguard. Slow review introduces different risks: habituation, cognitive offloading, institutionalized over-reliance, and gradual loss of independent verification capacity.

The practical implication is that automation bias should be treated as a design input rather than an afterthought. System design, process architecture, organizational accountability structures, and ongoing behavioral monitoring must all be structured to preserve active, skilled, and genuinely accountable human judgment in the review process. A human signature at the end of an AI-generated workflow is not the same as substantive human review of AI-assisted analysis. The existing NRC and IAEA human-

factors frameworks already supply the basis for this work; what they need is extension to deliberative-context automation bias, long-term behavioral monitoring, and team accountability design.

The objective is not to slow AI adoption but to direct it. In AI-assisted PRA and regulatory review, the most important question is not whether a human remains formally in the loop, but whether that human retains the skill, authority, accountability, and independence needed to detect error. Automation bias should therefore be treated as an emergent human-organizational risk factor: one that can be managed through design and governance, but only if it is recognized before over-reliance becomes routine.

REFERENCES

[1] K. Goddard, A. Roudsari, and J. C. Wyatt, *"Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators,"* Journal of the American Medical Informatics Association, vol. 19, no. 1, pp. 121–127, (2012), doi: 10.1136/amiajnl-2011-000089.

[2] R. Parasuraman and V. Riley, *"Humans and Automation: Use, Misuse, Disuse, Abuse,"* Human Factors, vol. 39, no. 2, pp. 230–253, (1997), doi: 10.1518/001872097778543886.

[3] L. J. Skitka, K. L. Mosier, and M. Burdick, *"Does Automation Bias Decision-Making?,"* International Journal of Human-Computer Studies, vol. 51, no. 5, pp. 991–1006, (1999), doi: 10.1006/ijhc.1999.0252.

[4] R. Parasuraman and D. H. Manzey, *"Complacency and Bias in Human Use of Automation: An Attentional Integration,"* Human Factors, vol. 52, no. 3, pp. 381–410, (2010), doi: 10.1177/0018720810376055.

[5] L. J. Skitka, K. L. Mosier, and M. Burdick, *"Accountability and Automation Bias,"* International Journal of Human-Computer Studies, vol. 52, no. 4, pp. 701–717, (2000), doi: 10.1006/ijhc.1999.0349.

[6] D. Lyell and E. Coiera, *"Automation Bias and Verification Complexity: A Systematic Review,"* Journal of the American Medical Informatics Association, vol. 24, no. 2, pp. 423–431, (2017), doi: 10.1093/jamia/ocw105.

[7] S. Jabbour *et al.*, *"Measuring the Impact of AI in the Diagnosis of Hospitalized Patients: A Randomized Clinical Vignette Survey Study,"* JAMA, vol. 330, no. 23, pp. 2275–2284, (2023), doi: 10.1001/jama.2023.22295.

[8] E. S. Berner *et al.*, *"Performance of Four Computer-Based Diagnostic Systems,"* New England Journal of Medicine, vol. 330, no. 25, pp. 1792–1796, (1994), doi: 10.1056/NEJM199406233302506.

[9] E. Goh *et al.*, *"Physician Clinical Decision Modification and Bias Assessment in a Randomized Controlled Trial of AI Assistance,"* Communications Medicine, vol. 5, p. 59, (2025), doi: 10.1038/s43856-025-00781-2.

[10] C. P. Friedman *et al.*, *"Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation: A Multisite Study of 2 Systems,"* JAMA, vol. 282, no. 19, pp. 1851–1856, (1999), doi: 10.1001/jama.282.19.1851.

[11] J. I. Westbrook, E. W. Coiera, and A. S. Gosling, *"Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions?,"* Journal of the American Medical Informatics Association, vol. 12, no. 3, pp. 315–321, (2005), doi: 10.1197/jamia.M1717.

[12] T. L. Tsai, D. B. Fridsma, and G. Gatti, *"Computer Decision Support as a Source of Interpretation Error: The Case of Electrocardiograms,"* Journal of the American Medical Informatics Association, vol. 10, no. 5, pp. 478–483, (2003), doi: 10.1197/jamia.M1279.

[13] R. Khera, M. A. Simon, and J. S. Ross, *"Automation Bias and Assistive AI: Risk of Harm From AI-Driven Clinical Decision Support,"* JAMA, vol. 330, no. 23, pp. 2255–2257, (2023), doi: 10.1001/jama.2023.22557.

[14] K. L. Mosier, L. J. Skitka, S. Heers, and M. Burdick, *"Automation Bias: Decision Making and Performance in High-Tech Cockpits,"* The International Journal of Aviation Psychology, vol. 8, no. 1, pp. 47–63, (1998), doi: 10.1207/s15327108ijap0801_3.

[15] Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, "Final Report on the Accident on 1st June 2009 to the Airbus A330-203 Registered F-GZCP Operated by Air France Flight AF

447 Rio de Janeiro – Paris," Bureau d'Enquêtes et d'Analyses pour la sécurité de l'aviation civile, Le Bourget Cedex, France, (2012). [Online]. Available: https://www.faa.gov/sites/faa.gov/files/AirFrance447_BEA.pdf
[16] J. K. Hawley, "Patriot Wars: Automation and the Patriot Air and Missile Defense System," (2017). Accessed: May 31, 2026. [Online]. Available: https://www.cnas.org/publications/reports/patriot-wars
[17] M. Kirchler *et al.*, *"The Effect of Fast and Slow Decisions on Risk Taking,"* Journal of Risk and Uncertainty, vol. 54, no. 1, pp. 37–59, (2017), doi: 10.1007/s11166-017-9252-4.
[18] M. G. Kocher, J. Pahlke, and S. T. Trautmann, *"Tempus Fugit: Time Pressure in Risky Decisions,"* Management Science, vol. 59, no. 10, pp. 2380–2391, (2013), doi: 10.1287/mnsc.2013.1711.
[19] J. M. O'Hara, J. C. Higgins, S. A. Fleger, and P. A. Pieringer, "Human Factors Engineering Program Review Model," U.S. Nuclear Regulatory Commission, Washington, DC, NUREG-0711, Revision 3, (2012). [Online]. Available: https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0711/r3/index
[20] J. M. O'Hara *et al.*, "Human Factors Considerations with Respect to Emerging Technology in Nuclear Power Plants," U.S. Nuclear Regulatory Commission, Washington, DC, NUREG/CR-6947, (2008).
[21] J. O'Hara, J. Higgins, and M. Pena, "Human-Performance Issues Related to the Design and Operation of Small Modular Reactors," U.S. Nuclear Regulatory Commission, Washington, DC, NUREG/CR-7126, (2012). [Online]. Available: https://www.nrc.gov/reading-rm/doc-collections/nuregs/contract/cr7126/index
[22] J. M. O'Hara and S. Fleger, "Human-System Interface Design Review Guidelines," U.S. Nuclear Regulatory Commission, Washington, DC, NUREG-0700, Revision 3, (2020). [Online]. Available: https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr0700/r3/index
[23] U.S. Nuclear Regulatory Commission, "Artificial Intelligence Strategic Plan: Fiscal Years 2023–2027," U.S. Nuclear Regulatory Commission, Washington, DC, NUREG-2261, (2023). [Online]. Available: https://www.nrc.gov/reading-rm/doc-collections/nuregs/staff/sr2261/index
[24] International Atomic Energy Agency, "Considerations for Deploying Artificial Intelligence Applications in the Nuclear Power Industry," International Atomic Energy Agency, Vienna, IAEA Nuclear Energy Series STI/PUB/2119, (2025). doi: 10.61092/iaea.s6uy-wjt8.
[25] International Atomic Energy Agency, "Regulatory Oversight of Human and Organizational Factors for Safety of Nuclear Installations," International Atomic Energy Agency, Vienna, IAEA-TECDOC-1846, (2018).